Lasso Screening Rules via Dual Polytope Projection

Jie Wang[1], Binbin Lin[2], Pinghua Gong[3], Peter Wonka[1], and Jieping Ye[1]

[1]Computer Science and Engineering, Arizona State University, USA
[2]State Key Lab of CAD & CG, Zhejiang University, P.R.China
[3]Department of Automation, Tsinghua University, P.R.China

November 19, 2012

Abstract

Lasso is a widely used regression technique to find sparse representations. When the dimension of the feature space and the number of samples are extremely large, solving the Lasso problem remains challenging. To improve the efficiency of solving large-scale Lasso problems, El Ghaoui and his colleagues have proposed the SAFE rules which are able to quickly identify the inactive predictors, i.e., predictors that have 0 components in the solution vector. Then, the inactive predictors or features can be removed from the optimization problem to reduce its scale. By transforming the standard Lasso to its dual form, it can be shown that the inactive predictors include the set of inactive constraints on the optimal dual solution. In this paper, we propose a fast and efficient screening rule via Dual Polytope Projections (DPP), which is mainly based on the uniqueness and nonexpansiveness of the optimal dual solution due to the fact that the feasible set in the dual space is a convex and closed polytope. Moreover, we show that our screening rule can be extended to identify inactive groups in group Lasso. To the best of our knowledge, there are currently no "exact" screening rules for group Lasso. We have evaluated our screening rule using both synthetic and real data sets. Results show that our rule is more effective to identify inactive predictors than existing state-of-the-art screening rules.

1 Introduction

Data with various structures and scales comes from almost every aspect of daily life. To effectively extract patterns in the data and build interpretable models with high prediction accuracy is always desirable. One popular technique to identify important explanatory features is by sparse regularization. For instance, consider the widely used ℓ_1-regularized least squares regression problem known as Lasso Tibshirani [1996]. The most appealing property of Lasso is the sparsity of the solutions, which is equivalent to feature selection. Suppose we have N observations and p predictors. Let $\mathbf{y}$ denote the N dimensional response vector and $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p]$ be the $N \times p$ feature matrix. The Lasso problem is formulated as the following optimization problem:

$$\inf_{\beta \in \Re^p} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1 \tag{1}$$

where $\lambda \geq 0$ is a regularization parameter.

Lasso has achieved great success in a wide range of applications Chen et al. [2001], Candès [2006], Zhao and Yu [2006], Bruckstein et al. [2009], Wright et al. [2010] and in recent years many algorithms have been developed to efficiently solve the Lasso problem Efron et al. [2004], Kim et al. [2007], Park and Hastie [2007], Donoho and Tsaig [2008], Friedman et al. [2007], Becker et al. [2010], Friedman et al. [2010]. However, when the dimension of feature space and the number of samples are very large, solving the Lasso problem remains challenging because we may not even be able to load the data matrix into main memory. The idea of a screening test proposed by El Ghaoui *et al.* El Ghaoui et al. [2010a] is to first identify inactive predictors

that have 0 components in the solution and then remove them from the optimization. Therefore, we can work on a reduced feature matrix to solve Lasso efficiently.

In El Ghaoui et al. [2010a], the "SAFE" rule discards $\mathbf{x}_i$ when

$$|\mathbf{x}_i^T\mathbf{y}| < \lambda - \|\mathbf{x}_i\|_2\|\mathbf{y}\|_2\frac{\lambda_{max}-\lambda}{\lambda_{max}} \tag{2}$$

where $\lambda_{max} = \max_i |\mathbf{x}_i^T\mathbf{y}|$ is the largest parameter such that the solution is nontrivial. Tibshirani et al. Tibshirani et al. [2012] proposed a set of strong rules which were more effective in identifying inactive predictors. The basic version discards $\mathbf{x}_i$ if

$$|\mathbf{x}_i^T\mathbf{y}| < 2\lambda - \lambda_{max} \tag{3}$$

However, it should be noted that the proposed strong rules might mistakenly discard active predictors, i.e., predictors which have nonzero coefficients in the solution vector. Xiang et al. Xiang et al. [2011], Xiang and Ramadge [2012] developed a set of screening tests based on the estimation of the optimal dual solution and they have shown that the SAFE rules are in fact a special case of the general sphere test.

In this paper, we develop new efficient and effective screening rules for the Lasso problem; our screening rules are exact in the sense that no active predictors will be discarded. By transforming problem (1) to its dual form, our motivation is mainly based on three geometric observations in the dual space. First, the active predictors belong to a subset of the active constraints on the optimal dual solution, which is a direct consequence of the KKT conditions. Second, the optimal dual solution is in fact the projection of the scaled response vector onto the feasible set of the dual variables. Third, because the feasible set of the dual variables is closed and convex, the projection is nonexpansive with respect to λ Bertsekas [2003], which results in an effective estimation of its variation.

The rest of this paper is organized as follows. We present the DPP screening rules for the Lasso problem in Section 2. Section 3 extends the idea of DPP screening rules to identify inactive groups in group Lasso Yuan and Lin [2006]. We have evaluated the proposed screening rules using both synthetic and real data. In Section 4, extensive experimental results demonstrate that the proposed rules are more effective than existing state-of-art screening rules.

2 Screening Rules for Lasso via Dual Polytope Projections

In this section, we first discuss the geometric properties of the dual formulation of problem (1) (Section 2.1). Specifically, the optimal dual solution can be formulated as the projection of the scaled response vector onto the feasible set, which is a closed and convex polytope in the dual space. According to the properties of projection operators with respect to closed convex sets Bertsekas [2003], the dual optimal is unique and nonexpansive. Based on the geometric properties of the dual optimal, we develop the fundamental principle, i.e., Theorem 1, which can be used to construct screening rules for Lasso. For illustrative purposes only, we provide Corollary 2 as a concrete example of the fundamental principle. We also reveal the connections between DPP rules and the sphere test Xiang et al. [2011]. In section 2.2, we discuss the relation between dual optimal and LARS Efron et al. [2004]. As a straightforward extension of DPP rules, we develop the sequential version of DPP (SDPP) in Section 2.3.

2.1 Fundamental Screening Rules via Dual Polytope Projections

Different from Xiang et al. [2011], Xiang and Ramadge [2012], we do not assume $\mathbf{y}$ and all $\mathbf{x}_i$ have unit length. We first transform problem (1) to its dual form (to make the paper self-contained, we provide the detailed derivation of the dual form in the supplemental materials):

$$\begin{aligned} \sup_{\theta} \quad & \frac{1}{2}\|\mathbf{y}\|_2^2 - \frac{\lambda^2}{2}\|\theta - \frac{\mathbf{y}}{\lambda}\|_2^2 \\ \text{subject to} \quad & |\mathbf{x}_i^T\theta| \leq 1,\ i = 1, 2, \ldots, p \end{aligned} \tag{4}$$

where θ is the dual variable.

Since the feasible set, denoted by F, is the intersection of $2p$ half-spaces, it is a closed and convex polytope. From the objective function of the dual problem (4), it is easy to see that the optimal dual solution θ^* is a feasible θ which is closest to $\frac{\mathbf{y}}{\lambda}$. In other words, θ^* is the projection of $\frac{\mathbf{y}}{\lambda}$ onto the polytope F. Mathematically, for an arbitrary vector $\mathbf{w}$ and a convex set C, if we define the projection function as

$$P_C(\mathbf{w}) = \underset{\mathbf{u}\in C}{\operatorname{argmin}} \|\mathbf{u} - \mathbf{w}\|_2,$$

then

$$\theta^* = P_F(\mathbf{y}/\lambda) = \underset{\theta\in F}{\operatorname{argmin}} \left\|\theta - \frac{\mathbf{y}}{\lambda}\right\|_2. \tag{5}$$

We know the optimal primal and dual solutions satisfy:

$$\mathbf{y} = \mathbf{X}\beta^* + \lambda\theta^* \tag{6}$$

and the KKT conditions for the Lasso problem (1) are

$$(\theta^*)^T\mathbf{x}_i \in \begin{cases} \text{sign}([\beta^*]_i) \text{ if } [\beta^*]_i \neq 0 \\ [-1, 1] \text{ if } [\beta^*]_i = 0 \end{cases} \tag{7}$$

where $[\cdot]_k$ denotes the k^{th} component.

By the KKT conditions in Eq. (7), if the inner product $(\theta^*)^T\mathbf{x}_i$ belongs to the open interval $(-1, 1)$, then the corresponding component $[\beta^*]_i$ in the solution vector $\beta^*(\lambda)$ has to be 0. As a result, $\mathbf{x}_i$ is an inactive predictor and can be removed from the optimization.

On the other hand, let

$$\partial H(\mathbf{x}_i) = \{\mathbf{z}\colon \mathbf{z}^T\mathbf{x}_i = 1\} \text{ and } H(\mathbf{x}_i)_- = \{\mathbf{z}\colon \mathbf{z}^T\mathbf{x}_i \leq 1\}$$

denote the hyperplane and half space determined by $\mathbf{x}_i$ respectively. Consider the dual problem (4); constraints induced by each $\mathbf{x}_i$ are equivalent to requiring each feasible θ to lie inside the intersection of $H(\mathbf{x}_i)_-$ and $H(-\mathbf{x}_i)_-$. If $|(\theta^*)^T\mathbf{x}_i| = 1$, i.e., either $\theta^* \in H(\mathbf{x}_i)_-$ or $\theta^* \in H(-\mathbf{x}_i)_-$, we say the constraints induced by $\mathbf{x}_i$ are active on θ^*.

We define the "*active*" set on θ^* as

$$\mathcal{I}_{\theta^*} = \{i\colon |(\theta^*)^T\mathbf{x}_i| = 1, i \in \mathcal{I}\}$$

where $\mathcal{I} = \{1, 2, \ldots, p\}$. Otherwise, if θ^* lies between $\partial H(\mathbf{x}_i)$ and $\partial H(-\mathbf{x}_i)$, i.e., $|(\theta^*)^T\mathbf{x}_i| < 1$, we can safely remove $\mathbf{x}_i$ from the problem because $[\beta^*]_i = 0$ according to the KKT conditions in Eq. (7). Similarly, the "*inactive*" set on θ^* is defined as $\overline{\mathcal{I}}_{\theta^*} = \mathcal{I} \setminus \mathcal{I}_{\theta^*}$.

Therefore, from a geometric perspective, if we know θ^*, i.e., the projection of $\frac{\mathbf{y}}{\lambda}$ onto F, the predictors in the inactive set on θ^* can be discarded from the optimization. It is worthwhile to mention that inactive predictors, i.e., predictors that have 0 components in the solution, are not the same as predictors in the inactive set. In fact, by the KKT conditions, predictors in the inactive set must be inactive predictors since they are guaranteed to have 0 components in the solution, but the converse may not be true.

Motivated by the above geometric intuitions, we next show how to find the predictors in the inactive set on θ^*. To emphasize the dependence on λ, let us write $\theta^*(\lambda)$ and $\beta^*(\lambda)$. If we know exactly where $\theta^*(\lambda)$ is, it will be trivial to find the predictors in the inactive set. Unfortunately, in most of the cases, we only have incomplete information about $\theta^*(\lambda)$ without actually solving problem (1) or (4). Suppose we know the exact $\theta^*(\lambda')$ for a specific λ'. How can we estimate $\theta^*(\lambda'')$ for another λ'' and its inactive set? To answer this question, we start from Eq. (5); $\theta^*(\lambda)$ is nonexpansive because it is a projection operator. We obtain the following result.

Theorem 1. *For the Lasso problem, assume we are given the solution of its dual problem $\theta^*(\lambda')$ for a specific λ'. Let λ'' be a nonnegative value different from λ'. If the following holds:*

$$|\mathbf{x}_i^T\theta^*(\lambda')| < 1 - \|\mathbf{x}_i\|_2\|\mathbf{y}\|_2\left|\frac{1}{\lambda'} - \frac{1}{\lambda''}\right|$$

then $[\beta^*(\lambda'')]_i = 0$.

Proof. From the KKT conditions in Eq. (7), we know

$$|\mathbf{x}_i^T\theta^*(\lambda'')| < 1 \Rightarrow [\beta^*(\lambda'')]_i = 0.$$

By the dual problem (4), $\theta^*(\lambda)$ is the projection of $\frac{\mathbf{y}}{\lambda}$ onto the feasible set F. According to the projection theorem Bertsekas [2003] for closed convex sets, $\theta^*(\lambda)$ is continuous and nonexpansive, i.e.,

$$\|\theta^*(\lambda'') - \theta^*(\lambda')\|_2 \leq \left\|\frac{\mathbf{y}}{\lambda''} - \frac{\mathbf{y}}{\lambda'}\right\|_2 = \|\mathbf{y}\|_2\left|\frac{1}{\lambda''} - \frac{1}{\lambda'}\right| \tag{8}$$

Then

$$\begin{aligned}
|\mathbf{x}_i^T\theta^*(\lambda'')| &\leq |\mathbf{x}_i^T\theta^*(\lambda'') - \mathbf{x}_i^T\theta^*(\lambda')| \\
&\quad + |\mathbf{x}_i^T\theta^*(\lambda')| \\
&< \|\mathbf{x}_i\|_2\|(\theta^*(\lambda'') - \theta^*(\lambda'))\|_2 \\
&\quad + 1 - \|\mathbf{x}_i\|_2\|\mathbf{y}\|_2\left|\frac{1}{\lambda''} - \frac{1}{\lambda'}\right| \\
&\leq \|\mathbf{x}_i\|_2\|\mathbf{y}\|_2\left|\frac{1}{\lambda''} - \frac{1}{\lambda\prime}\right| \\
&\quad + 1 - \|\mathbf{x}_i\|_2\|\mathbf{y}\|_2\left|\frac{1}{\lambda''} - \frac{1}{\lambda'}\right| \\
&= 1
\end{aligned} \tag{9}$$

which completes the proof. □

From theorem 1, it is easy to see our rule is quite flexible since every $\theta^*(\lambda')$ would result in a new screening rule. And the smaller the gap between λ' and λ'', the more effective the screening rule is. By "*more effective*", we mean a stronger capability of the screening rule in identifying inactive predictors.

As an example, let us find out $\theta^*(\lambda_{max})$. Recall that $\lambda_{max} = \max_i |\mathbf{x}_i^T\mathbf{y}|$. It is easy to verify $\frac{\mathbf{y}}{\lambda_{max}}$ is itself feasible. Therefore the projection of $\frac{\mathbf{y}}{\lambda_{max}}$ onto F is itself, i.e., $\theta^*(\lambda_{max}) = \frac{\mathbf{y}}{\lambda_{max}}$. Moreover, by noting that for $\forall\lambda > \lambda_{max}$, we have $|\mathbf{x}_i^T\mathbf{y}| < 1, i \in \mathcal{I}$, i.e., all predictors are in the inactive set at $\theta^*(\lambda)$, we conclude that the solution to problem (1) is 0. Combining all these together and plugging $\theta^*(\lambda_{max}) = \frac{\mathbf{y}}{\lambda_{max}}$ into Eq. (14), we obtain the following screening rule.

Corollary 2. DPP*: For the Lasso problem (1), let $\lambda_{max} = \max_i |\mathbf{x}_i^T\mathbf{y}|$.*

1. *If $\lambda > \lambda_{max}$, then $[\beta^*]_i = 0, \forall i \in \mathcal{I}$;*
2. *Otherwise, if the following holds:*

$$\left|\mathbf{x}_i^T\frac{\mathbf{y}}{\lambda_{max}}\right| < 1 - \|\mathbf{x}_i\|_2\|\mathbf{y}\|_2\left(\frac{1}{\lambda} - \frac{1}{\lambda_{max}}\right)$$

then $[\beta^*(\lambda)]_i = 0$.

Table 1: Illustration of the running time for DPP screening and for solving the Lasso problem after screening. T_s: time for screening. T_l: time for solving the Lasso problem after screening. T_o: the total time. Entries of the response vector $\mathbf{y}$ are i.i.d. by a standard Gaussian. Columns of the data matrix $\mathbf{X} \in \Re^{1000\times 100000}$ are generated by $\mathbf{x}_i = \mathbf{y} + \alpha\mathbf{z}$ where α is a random number drawn from the uniform distribution in $[0, 1]$. Entries of $\mathbf{z}$ are i.i.d. by a standard Gaussian. $\lambda_{max} = 0.95$ and λ/λ_{max}=0.5.

	Lasso	DPP	DPP2	DPP5	DPP10	DPP20
T_s (s)	—	0.035	0.073	0.152	0.321	0.648
T_l (s)	—	10.250	9.634	8.399	1.369	0.121
T_o (s)	103.314	10.285	9.707	8.552	1.690	0.769

Clearly, DPP is most effective when λ is close to λ_{max}. So how can we find a new $\theta^*(\lambda')$ with $\lambda' < \lambda_{max}$? Note that Eq. (6) is in fact a natural bridge which relates the primal and dual optimal solutions. As long as we know $\beta^*(\lambda')$, it is easy to get $\theta^*(\lambda')$ when λ is relatively small, e.g., LARS Efron et al. [2004] and Homotopy Osborne et al. [2000] algorithms.

Remark: Xiang *et al.* Xiang et al. [2011] developed a general sphere test which says that if θ^* is estimated to be inside a ball $\|\theta^* - \mathbf{q}\|_2 \leq r$, then

$$|\mathbf{x}_i^T\mathbf{q}| < (1 - r) \Rightarrow [\beta^*]_i = 0.$$

Considering the DPP rules in Theorem 1, it is equivalent to setting $\mathbf{q} = \theta^*(\lambda')$ and $r = |\frac{1}{\lambda'} - \frac{1}{\lambda''}|$. Therefore, different from the sphere test and Dome developed in Xiang et al. [2011], Xiang and Ramadge [2012] with the radius r fixed at the beginning, the construction of our DPP rules is equivalent to an "r" decreasing process. Clearly, the smaller r is, the more inactive predictors we can discard and the more effective the DPP rules will be.

Remark: It is worthwhile to note that DPP is not the same as ST1 in Xiang et al. [2011] and SAFE in El Ghaoui et al. [2010a]. From the perspective of the sphere test, the radius of ST1/SAFE and DPP are the same. But the centers of ST1 and DPP are x/λ and x/λ_{max} respectively, which leads to different formulas, i.e., Eq. (2) and Corollary 2.

2.2 DPP Rules with LARS/Homotopy Algorithms

It is well known that under mild conditions, the set $\{\beta^*(\lambda) : \lambda > 0\}$ (also know as regularization path Mairal and Yu [2012]) is continuous piecewise linear Osborne et al. [2000], Efron et al. [2004], Mairal and Yu [2012]. The output of LARS or Homotopy algorithms is in fact a sequence of values like $(\beta^*(\lambda^{(0)}), \lambda^{(0)}), (\beta^*(\lambda^{(1)}), \lambda^{(1)}), \ldots$, where $\beta^*(\lambda^{(i)})$ corresponds to the ith breakpoint of the regularization path $\{\beta^*(\lambda) : \lambda > 0\}$ and $\lambda^{(i)}$s are monotonically decreasing. By Eq. (6), once we get $\beta^*(\lambda^{(i)})$, we can immediately compute $\theta^*(\lambda^{(i)})$. Then according to Theorem 1, we can construct a DPP rule based on $\theta^*(\lambda^{(i)})$ and $\lambda^{(i)}$. For convenience, if the DPP rule is built based on $\theta^*(\lambda^{(i)})$, we add the index i as suffix to DPP, e.g., DPP5 means it is developed based on $\theta^*(\lambda^{(5)})$.

It should be noted that LARS or Homotopy algorithms are very efficient to find the first few breakpoints of the regularization path and the corresponding parameters. For the first few breakpoints, the computational cost is roughly $O(Np)$, i.e., linear with the size of the data matrix $\mathbf{X}$. In Table 1, we report both the time used for screening and the time needed to solve the Lasso problem after screening. The Lasso solver is from the SLEP Liu et al. [2009] package.

From Table 1, we can see that compared with the time saved by the screening rules, the time used for screening is negligible. The efficiency of the Lasso solver is improved by DPP20 more than 130 times. In practice, DPP rules built on the first few $\theta^*(\lambda^{(i)})$'s lead to more significant performance improvement than existing state-of-art screening tests. We will demonstrate the effectiveness of our DPP rules in the experiment section.

As another useful property of LARS/Homotopy algorithms, it is worthwhile to mention that changes of the active set only happen at the breakpoints Osborne et al. [2000], Efron et al. [2004], Mairal and

Yu [2012]. Consequently, given the parameters corresponding to a pair of adjacent breakpoints, e.g., $\lambda^{(i)}$ and $\lambda^{(i+1)}$, the active set for $\lambda \in (\lambda^{(i+1)}, \lambda^{(i)})$ is the same as $\lambda = \lambda^{(i)}$. Therefore, besides the sequence of breakpoints and the associated parameters $(\beta^*(\lambda^{(0)}), \lambda^{(0)}), \ldots (\beta^*(\lambda^{(k)}), \lambda^{(k)})$ computed by LARS/Homotopy algorithms, we know the active set for $\forall \lambda \geq \lambda^{(k)}$. Hence we can remove the predictors in the inactive set from the optimization problem (1). This scheme has been embedded in DPP rules.

Remark: Some works, e.g., Tibshirani et al. [2012] and El Ghaoui et al. [2010b], solve several Lasso problems for different parameters to improve the screening performance. However, the DPP algorithms do not aim to solve a sequence of Lasso problems, but just to accelerate one. The LARS/Homotopy algorithms are used to find the first few breakpoints of the regularization path and the corresponding parameters, instead of solving general Lasso problems. Therefore, different from Tibshirani et al. [2012] and El Ghaoui et al. [2010b] who need to iteratively compute a screening step and a Lasso step, DPP algorithms only compute one screening step and one Lasso step.

2.3 Sequential Version of DPP Rules

Motivated by the ideas of Tibshirani et al. [2012] and El Ghaoui et al. [2010b], we can develop a sequential version of DPP rules. In other words, if we are given a sequence of parameter values $\lambda_1 > \lambda_2 > \ldots > \lambda_m$, we can first apply DPP to discard inactive predictors for the Lasso problem (1) with parameter being λ_1. After solving the reduced optimization problem for λ_1, we obtain the exact solution $\beta^*(\lambda_1)$. Hence by Eq. (6), we can find $\theta^*(\lambda_1)$. According to Theorem 1, once we know the optimal dual solution $\theta^*(\lambda_1)$, we can construct a new screening rule to identify inactive predictors for problem (1) with $\lambda = \lambda_2$. By repeating the above process, we obtain the sequential version of the DPP rule (SDPP).

Corollary 3. SDPP: *For the Lasso problem (1), suppose we are given a sequence of parameter values $\lambda_{max} = \lambda_0 > \lambda_1 > \ldots > \lambda_m$. Then for any integer $0 \leq k < m$, if $\beta^*(\lambda_k)$ is known and the following holds:*

$$\left| \mathbf{x}_i^T \frac{\mathbf{y} - \mathbf{X}\beta^*(\lambda_k)}{\lambda_k} \right| < 1 - \|\mathbf{x}_i\|_2 \|\mathbf{y}\|_2 \big(\frac{1}{\lambda_{k+1}} - \frac{1}{\lambda_k} \big)$$

then $[\beta^*(\lambda_{k+1})]_i = 0$.

Remark: There are some other related works on screening rules, e.g., Wu *et al.* Wu et al. [2009] built screening rules for l_1 penalized logistic regression based on the inner products between the response vector and each predictor; Tibshirani *et al.* Tibshirani et al. [2012] developed strong rules for a set of Lasso-type problems via the inner products between the residual and predictors; in Fan and Lv [2008], Fan and Lv studied screening rules for Lasso and related problems. But all of the above works may mistakenly discard predictors that have non-zero coefficients in the solution. Similar to El Ghaoui et al. [2010a], Xiang et al. [2011], Xiang and Ramadge [2012], our DPP rules are exact in the sense that the predictors discarded by our rules are inactive predictors, i.e., predictors that have zero coefficients in the solution.

3 Extensions to Group Lasso

To demonstrate the flexibility of DPP rules, we extend our idea to the group Lasso problem Yuan and Lin [2006]:

$$\inf_{\beta \in \Re^p} \frac{1}{2} \|\mathbf{y} - \sum_{g=1}^{G} \mathbf{X}_g \beta_g\|_2^2 + \lambda \sum_{g=1}^{G} \sqrt{n_g} \|\beta_g\|_2 \tag{10}$$

where $\mathbf{X}_g \in \Re^{N \times n_g}$ is the data matrix for the gth group and $p = \sum_{g=1}^{G} n_g$.

The corresponding dual problem of (10) is (see detailed derivation in the supplemental materials):

$$\begin{aligned} \sup_{\theta} \quad & \frac{1}{2}\|\mathbf{y}\|_2^2 - \frac{\lambda^2}{2}\|\theta - \frac{\mathbf{y}}{\lambda}\|_2^2 \\ \text{subject to} \quad & \|\mathbf{X}_g^T \theta\|_2 \leq \sqrt{n_g},\ g = 1, 2, \ldots, G \end{aligned} \tag{11}$$

Similar to the Lasso problem, the primal and dual optimal solutions of the group Lasso satisfy:

$$\mathbf{y} = \sum_{g=1}^{G} \mathbf{X}_g \beta_g^* + \lambda \theta^* \tag{12}$$

and the KKT conditions are:

$$(\theta^*)^T \mathbf{X}_g \in \begin{cases} \sqrt{n_g} \frac{\beta_g^*}{\|\beta_g^*\|_2} \text{ if } \beta_g^* \neq 0 \\ \sqrt{n_g}\mathbf{u}, \|\mathbf{u}\|_2 \leq 1 \text{ if } \beta_g^* = 0 \end{cases} \tag{13}$$

for $g = 1, 2, \ldots, G$.

Clearly, if $\|(\theta^*)^T \mathbf{X}_g\|_2 < \sqrt{n_g}$, we can conclude that $\beta_g^* = 0$.

Consider problem (11). It is easy to see that the dual optimal θ^* is the projection of $\frac{\mathbf{y}}{\lambda}$ onto the feasible set. For each g, the constraint $\|\mathbf{X}_g^T \theta\|_2 \leq \sqrt{n_g}$ confines θ to an ellipsoid which is closed and convex. Therefore, the feasible set of the dual problem (11) is the intersection of ellipsoids and thus closed and convex. Hence $\theta^*(\lambda)$ is also nonexpansive for the group lasso problem. Similar to Theorem 1, we can readily develop the following theorem for group Lasso.

Theorem 4. *For the group Lasso problem, assume we are given the solution of its dual problem $\theta^*(\lambda')$ for a specific λ'. Let λ'' be a nonnegative value different from λ'. If the following holds:*

$$\|\mathbf{X}_g^T \theta^*(\lambda')\|_2 < \sqrt{n_g} - \|\mathbf{X}_g\|_F \|\mathbf{y}\|_2 \left| \frac{1}{\lambda'} - \frac{1}{\lambda''} \right| \tag{14}$$

then $\beta_g^*(\lambda'') = 0$.

Proof. From the KKT conditions in Eq. (13), we know

$$\|\mathbf{X}_g^T \theta^*(\lambda'')\|_2 < \sqrt{n_g} \Rightarrow \beta_g^*(\lambda'') = 0.$$

By the dual problem (11), $\theta^*(\lambda)$ is the projection of $\frac{\mathbf{y}}{\lambda}$ onto the feasible set which is closed and convex. Note, the feasible set is in fact the intersection of ellipsoids:

$$\{\theta \colon \|\mathbf{X}_g^T \theta\|_2 \leq \sqrt{n_g}\},\ g = 1, 2, \ldots, G.$$

According to the projection theorem Bertsekas [2003] for closed convex sets, $\theta^*(\lambda)$ is continuous and nonexpansive, i.e.,

$$\|\theta^*(\lambda'') - \theta^*(\lambda')\|_2 \leq \left\| \frac{\mathbf{y}}{\lambda''} - \frac{\mathbf{y}}{\lambda'} \right\|_2 = \|\mathbf{y}\|_2 \left| \frac{1}{\lambda''} - \frac{1}{\lambda'} \right| \tag{15}$$

Then

$$\begin{aligned} \|\mathbf{X}_g^T \theta^*(\lambda'')\|_2 &\leq \|\mathbf{X}_g^T \theta^*(\lambda'') - \mathbf{X}_g^T \theta^*(\lambda')\|_2 \\ &\quad + \|\mathbf{X}_g^T \theta^*(\lambda')\|_2 \\ &< \|\mathbf{X}_g\|_2 \|(\theta^*(\lambda'') - \theta^*(\lambda'))\|_2 \\ &\quad + \sqrt{n_g} - \|\mathbf{X}_g\|_F \|\mathbf{y}\|_2 \left| \frac{1}{\lambda'} - \frac{1}{\lambda''} \right| \\ &\leq \|\mathbf{X}_g\|_F \|\mathbf{y}\|_2 \left| \frac{1}{\lambda''} - \frac{1}{\lambda\prime} \right| \\ &\quad + \sqrt{n_g} - \|\mathbf{X}_g\|_F \|\mathbf{y}\|_2 \left| \frac{1}{\lambda'} - \frac{1}{\lambda''} \right| \\ &= \sqrt{n_g} \end{aligned} \tag{16}$$

which completes the proof.

We use the fact that $\|\mathbf{X}_g\|_2 \leq \|\mathbf{X}_g\|_F$ in the last inequality of Eq. (16). The subscript $\|\cdot\|_F$ denotes the Frobenius norm. □

Similar to the Lasso problem, let

$$\lambda_{max} = \max_g \frac{\|\mathbf{X}_g^T \mathbf{y}\|_2}{\sqrt{n_g}};$$

it is easy to see that $\frac{\mathbf{y}}{\lambda_{max}}$ is itself feasible, and λ_{max} is the largest parameter such that problem (10) has a nonzero solution. Similar to DPP and SDPP, we can construct GDPP and SGDPP for group Lasso.

Corollary 5. GDPP: *For the group Lasso problem (10), let* $\lambda_{max} = \max_g \frac{\|\mathbf{X}_g^T \mathbf{y}\|_2}{\sqrt{n_g}}$.

1. *If* $\lambda > \lambda_{max}$, $\beta_g^*(\lambda) = 0, \forall g = 1, 2, \ldots, G$;
2. *Otherwise, if the following holds:*

$$\left| \mathbf{X}_g^T \frac{\mathbf{y}}{\lambda_{max}} \right| < \sqrt{n_g} - \|\mathbf{X}_g\|_F \|\mathbf{y}\|_2 \big(\frac{1}{\lambda} - \frac{1}{\lambda_{max}}\big)$$

then $\beta_g^*(\lambda) = 0$.

Corollary 6. SGDPP: *For the group Lasso problem (10), suppose we are given a sequence of parameter values* $\lambda_{max} = \lambda_0 > \lambda_1 > \ldots > \lambda_m$. *For any integer* $0 \leq k < m$, *if* $\beta^*(\lambda_k)$ *is known and the following holds:*

$$\left| \mathbf{X}_g^T \frac{\mathbf{y} - \sum_{g=1}^G \mathbf{X}_g \beta_g^*(\lambda_k)}{\lambda_k} \right| < \sqrt{n_g} - \|\mathbf{X}_g\|_F \|\mathbf{y}\|_2 \big(\frac{1}{\lambda_{k+1}} - \frac{1}{\lambda_k}\big)$$

then $\beta_g^*(\lambda_{k+1}) = 0$.

4 Experiments

We evaluated our screening rules on both synthetic and real data sets. To measure the performance of our screening rules, we compute the rejection rate, i.e., the ratio between the number of predictors discarded by screening rules and the actual number of zero predictors in the ground truth. Because the DPP rules are exact, i.e., no active predictors will be mistakenly discarded, the rejection rate will be less than one.

We compare the performance of DPP with Dome Xiang and Ramadge [2012], Xiang et al. [2011] which achieves state-of-art performance for the Lasso problem among exact screening rules Xiang and Ramadge [2012]. We evaluate GDPP and SGDPP for the group Lasso problem on three synthetic data sets in section 4.2. We are not aware of any "*exact*" screening rules for the group Lasso problem at this point. For SAFE and Dome, it is not straightforward to extend them to the group Lasso problem.

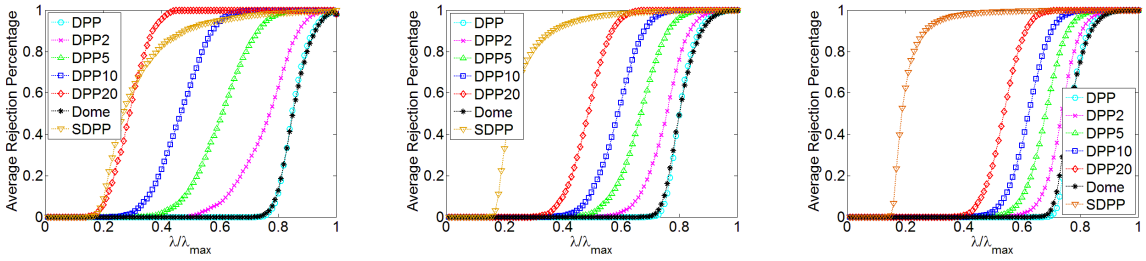
Similarly to previous works Xiang et al. [2011], we do not report the computational time saved by screening because it can be easily computed from the projection ratio. Specifically, if the Lasso solver is linear in terms of the size of the data matrix $\mathbf{X}$, a $K\%$ rejection of the data can save $K\%$ computational time.

4.1 DPPs for the Lasso Problem

We compare the performance of DPP rules and Dome on: (**a**) three synthetic datasets with different dimensions; (**b**) the MNIST handwritten digit data set Lecun et al. [1998]; (**c**) the COIL rotational image data set Nene et al. [1996], and (**d**) the Olivetti Faces data set Samaria and Harter [1994]. There are many solvers Becker et al. [2010], Friedman et al. [2007], Kim et al. [2007], Osborne et al. [2000] which can be used to find the ground truth, i.e., the solution of problem (1).

4.1.1 Synthetic Data Sets

We generate three synthetic data sets with different dimensions. For each of the cases, the entries of data matrix $\mathbf{X}$ and response vector $\mathbf{y}$ are independent identically distributed by a standard Gaussian. Each data matrix contains 100 samples with $p = 50, 500,$ and 5000 respectively. For each case, once we generate the data matrix $\mathbf{X}$, we compare the performance of DPP rules with Dome along a sequence of 100 parameter values equally spaced on the λ/λ_{max} scale. Then we repeat the procedure 500 times and report the average performance of each rule.



(a) $\mathbf{X} \in \Re^{100\times 50}$ (b) $\mathbf{X} \in \Re^{100\times 500}$ (c) $\mathbf{X} \in \Re^{100\times 5000}$

Figure 1: Comparison of DPP rules and Dome on three synthetic datasets. Each column corresponds to each of the three synthetic data sets with different dimensions.

The three subfigures of Fig. 1 correspond to the three different design matrices $\mathbf{X}$ and the average λ_{max} is 0.249, 0.315 and 0.371 respectively. As shown in Fig. 1, the performance of DPP is comparable to Dome but all the other DPP rules significantly outperform Dome. In contrast to Dome which performs better with larger λ_{max} Xiang et al. [2011], DPP rules exhibit stronger capability in discarding inactive predictors when λ_{max} is small. The geometric intuition behind this observation is due to the fact that the sparser the predictors distribute over the unit ball, the longer the line segment of the regularization path is. If the length of the line segment of the regularization path is larger, the first few breakpoints may correspond to very small λ values.

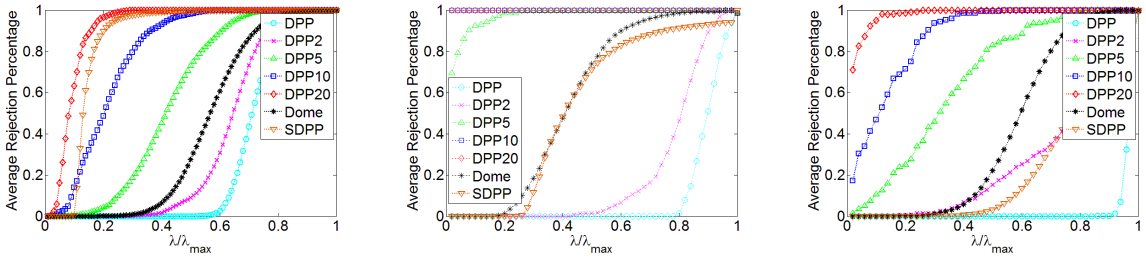
4.1.2 MNIST Digit Data Set

This data set contains grey images of scanned handwritten digits, including $60,000$ for training and $10,000$ for testing. The dimension of each image is 28×28. We first randomly select 100 images for each digit (and in total we have 1000 images) and get a data matrix $\mathbf{X} \in \Re^{784\times 1000}$.

Similarly, we compare the performance of DPP rules with Dome along a sequence of 100 parameter values equally spaced on the λ/λ_{max} scale. We repeat the procedure 500 times and report the average performance of each rule. In contrast to the case of synthetic data, the average λ_{max} is large ($\lambda_{max} = 0.837$) for the MNIST data set. As noted in Xiang et al. [2011], Xiang and Ramadge [2012], Dome is strong when λ_{max} is large. Fig. 2(a) shows Dome outperforms DPP and DPP2. But still, all the other DPP rules perform significantly better than Dome.

4.1.3 COIL Rotational Object Image Data Set

In this experiment, we consider the case where $N \gg p$ and the predictors are highly correlated. The COIL data set includes 7200 images for 100 objects. We use object No. 13 with 72 color images of size 128×128 taken every 5 degree by rotating the object. Each time, we take one of the images as the response vector $\mathbf{y}$ and use all the remaining images to construct the data matrix. Then we compare the performance of DPP rules and Dome along a sequence of 50 parameter values equally spaced on the λ/λ_{max} scale. By using every image as response vector, we repeat the procedure 72 times. We transform each color image to a column



(a) MNIST (b) COIL (c) Olivetti

Figure 2: Comparison of DPP rules and Dome on three real datasets. MNIST digit data set (left), COIL image data set (middle) and Olivetti face data set (right).

vector with $3 \times 128 \times 128 = 49152$ elements. Therefore we obtain a data matrix $\mathbf{X} \in \Re^{49152 \times 71}$. The average λ_{max} is 0.988.

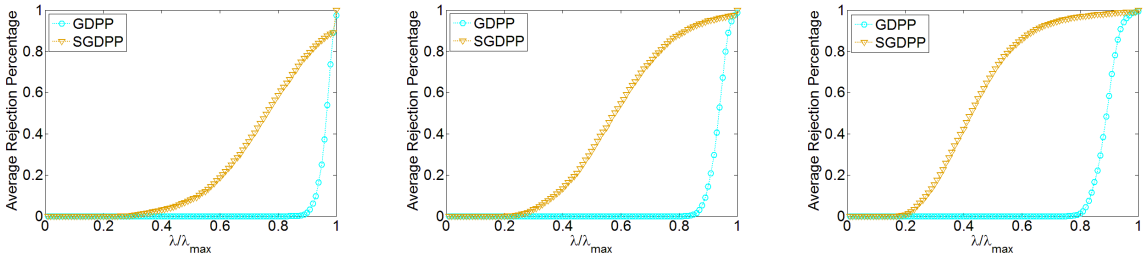
As shown in Fig. 2(b), Dome discards much more inactive predictors than DPP and DPP2 even for small λ. But DPP5 significantly outperforms Dome. This is because the average parameter value of the 5^{th} breakpoint $\overline{\lambda}^{(5)}$ is very small, and we know for any single run DPP5 can discard all predictors in the inactive set for $\lambda \geq \lambda^{(5)}$. For the same reason, DPP10 and DPP20 can identify almost all of the inactive predictors even for very small λ.

4.1.4 Olivetti Faces Data Set

This data set includes 400 grey scale face images of size 64×64 for 40 people (10 for each). We sequentially take one of the images as response vectors and the left images to construct data matrix $\mathbf{X}$. All images are converted to column vectors and thus $\mathbf{y} \in \Re^{4096}$, $\mathbf{X} \in \Re^{4096 \times 399}$. We compare the performance of DPP rules and Dome along a sequence of 50 parameter values equally spaced on the λ/λ_{max} scale. The average λ_{max} is 0.989.

As shown in Fig. 2(c), Dome outperforms DPP and DPP2. DPP5 discards more inactive predictors than Dome, especially for small λ. As expected, DPP10 and DPP20 further improve DPP5.

4.2 GDPPs for the Group Lasso Problem



(a) 20 groups (b) 50 groups (c) 100 groups

Figure 3: Performance of GDPP and SGDPP applied to three synthetic data sets with different number of groups.

We apply GDPPs to three synthetic data sets. The entries of data matrix $\mathbf{X} \in \Re^{100 \times 1000}$ and the response vector $\mathbf{y}$ are generated i.i.d. from the standard Gaussian distribution. For each of the cases, we randomly

divided $\mathbf{X}$ into 20, 50, and 100 groups. We compare the performance of GDPP and SGDPP along a sequence of 100 parameter values equally spaced on the λ/λ_{max} scale. We repeat the above procedure 100 times for each of the cases and report the average performance. The average λ_{max} values are 0.136, 0.167, and 0.219 respectively.

As shown in Fig. 3, it is expected that SGDPP significantly outperforms GDPP which only makes use of the information of the dual optimal solution at a single point.

Remark: For the group Lasso problem, the feasible set of its dual variables is the intersection of ellipsoids and is thus no longer a polytope. As a consequence, the path of the optimal solution is no longer piecewise linear. Due to this fact, it is more complicated to characterize the path and find the breakpoints where groups of predictors enter or leave the active set. However, if there are efficient algorithms which can find the breakpoints and the corresponding parameters like LARS for Lasso, we can potentially make use of those breakpoints and the associated parameters to construct more effective screening rules based on Theorem 4.

5 Conclusion

In this paper, we develop new screening rules for the Lasso problem by making use of the nonexpansiveness of the projection operator with respect to a closed convex set. Our new methods, i.e., DPP screening rules, are able to effectively identify inactive predictors of the Lasso problem, thus greatly reducing the size of the optimization problem. The idea of DPP rules can be easily generalized to screen the inactive groups of the group Lasso problem. Extensive numerical experiments on both synthetic and real data demonstrate the effectiveness of the proposed rules. It is worthwhile to mention that DPP rules can be combined with any Lasso solver as a speedup tool.

In the future, we plan to generalize our idea to other sparse formulations consisting of different loss functions, e.g., logistic/hinge loss, and more general structured sparse penalty, e.g., group/graph Lasso.

References

S. R. Becker, E. Candès, and M. Grant. Templates for convex cone problems with applications to sparse signal recovery. Technical report, Standford University, 2010.

D. P. Bertsekas. *Convex Analysis and Optimization*. Athena Scientific, 2003.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

A. Bruckstein, D. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review*, 51:34–81, 2009.

E. Candès. Compressive sampling. In *Proceedings of the International Congress of Mathematics*, 2006.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43: 129–159, 2001.

D. L. Donoho and Y. Tsaig. Fast solution of l-1 norm minimization problems when the solution may be sparse. *IEEE Trans. Inform. Theory*, 54:4789–4812, 2008.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Statist.*, 32:407–499, 2004.

L. El Ghaoui, V. Viallon, and T. Rabbani. Safe feature elimination in sparse supervised learning. Technical Report UC/EECS-2010-126, EECS Dept., University of California at Berkeley, September 2010a.

L. El Ghaoui, V. Viallon, and T. Rabbani. Safe feature elimination for the lasso and sparse supervised learning problems. arXiv:1009.4219v2, 2010b.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature spaces. *J. R. Statist. Soc. B*, 70:849–911, 2008.

J. Friedman, T. Hastie, H. Hëfling, and R. Tibshirani. Pathwise coordinate optimization. *Ann. Appl. Statis*, 1:302–332, 2007.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22, 2010.

S. J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. An interior-point method for large scale l1-regularized least squares. *IEEE J. Select. Top. Sign. Process.*, 1:606–617, 2007.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, 1998.

J. Liu, S. Ji, and J. Ye. *SLEP: Sparse Learning with Efficient Projections*. Arizona State University, 2009. URL `http://www.public.asu.edu/~jye02/Software/SLEP`.

J. Mairal and B. Yu. Complexity analysis of the lasso regularization path. In *ICML*, 2012.

S. A. Nene, S. K. Nayar, and H. Murase. Columbia object image library (coil). Technical report, No. CUCS-006-96, Dept. Comp. Science, Columbia University, 1996.

M. R. Osborne, B. Presnell, and B. A. Turlach. A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20:389–404, 2000.

M. Y. Park and T. Hastie. L1-regularized path algorithm for generalized linear models. *J. R. Statist. Soc. B*, 69:659–677, 2007.

F. Samaria and A. Harter. Parameterisation of a stochastic model for human face identification. In *Proceedings of 2nd IEEE Workshop on Applications of Computer Vision*, 1994.

R. Tibshirani. Regression shringkage and selection via the lasso. *J. R. Statist. Soc. B*, 58:267–288, 1996.

R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. Tibshirani. Strong rules for discarding predictors in lasso-type problems. *J. R. Statist. Soc. B*, 74:245–266, 2012.

J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, and S. Yan. Sparse representation for computer vision and pattern recognition. In *Proceedings of IEEE*, 2010.

T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange. Genomewide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25:714–721, 2009.

Z. J. Xiang and P. J. Ramadge. Fast lasso screening tests based on correlations. In *IEEE ICASSP*, 2012.

Z. J. Xiang, H. Xu, and P. J. Ramadge. Learning sparse representation of high dimensional data on large scale dictionaries. In *NIPS*, 2011.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B*, 68:49–67, 2006.

P. Zhao and B. Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006.

Appendix

In this appendix, we will show the detailed procedure to derive the dual formulation of standard lasso and group lasso in sections A and B.

A Deviation of the Dual Problem of Standard Lasso

A.1 Dual Formulation

Assuming the data matrix is $\mathbf{X} \in \Re^{N \times p}$, the standard Lasso problem is given by:

$$\inf_{\beta \in \Re^p} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1 \tag{17}$$

For completeness, we give a detailed deviation of the dual formulation of (17) in this section. Note that problem (17) has no constraints. Therefore the dual problem is trivial and useless. A common trick [Boyd and Vandenberghe, 2004] is to introduce a new set of variables $\mathbf{z} = \mathbf{y} - \mathbf{X}\beta$ such that problem (17) becomes:

$$\begin{aligned} \inf_{\beta} \quad & \frac{1}{2}\|\mathbf{z}\|_2^2 + \lambda\|\beta\|_1 \\ \text{subject to} \quad & \mathbf{z} = \mathbf{y} - \mathbf{X}\beta \end{aligned} \tag{18}$$

By introducing the dual variables $\eta \in \Re^N$, we get the Lagrangian of problem (18):

$$L(\beta, \mathbf{z}, \eta) = \frac{1}{2}\|\mathbf{z}\|_2^2 + \lambda\|\beta\|_1 + \eta^T \cdot (\mathbf{y} - \mathbf{X}\beta - \mathbf{z}) \tag{19}$$

For the Lagrangian, the primal variables are β and $\mathbf{z}$. And the dual function $g(\eta)$ is:

$$g(\eta) = \inf_{\beta, \mathbf{z}} L(\beta, \mathbf{z}, \eta) = \eta^T\mathbf{y} + \inf_{\beta}(-\eta^T\mathbf{X}\beta + \lambda\|\beta\|_1) + \inf_{\mathbf{z}}\left(\frac{1}{2}\|\mathbf{z}\|_2^2 - \eta^T\mathbf{z}\right) \tag{20}$$

In order to get $g(\eta)$, we need to solve the following two optimization problems.

$$\inf_{\beta} -\eta^T\mathbf{X}\beta + \lambda\|\beta\|_1 \tag{21}$$

and

$$\inf_{\mathbf{z}} \frac{1}{2}\|\mathbf{z}\|_2^2 - \eta^T\mathbf{z} \tag{22}$$

Let us first consider problem (21). Denote the objective function of problem (21) as

$$f_1(\beta) = -\eta^T\mathbf{X}\beta + \lambda\|\beta\|_1. \tag{23}$$

$f_1(\beta)$ is convex but not smooth. Therefore let us consider its subgradient

$$\partial f_1(\beta) = -\mathbf{X}^T\eta + \lambda\mathbf{v}$$

in which $\|\mathbf{v}\|_\infty \leq 1$ and $\mathbf{v}^\mathbf{T}\beta = \|\beta\|_1$, i.e., $\mathbf{v}$ is the subgradient of $\|\beta\|_1$.

The necessary condition for f_1 to attain an optimum is

$$\exists\, \beta', \text{ such that } 0 \in \partial f_1(\beta') = \{-\mathbf{X}^T\eta + \lambda\mathbf{v}'\}$$

where $\mathbf{v}' \in \partial\|\beta'\|_1$. In other words, $\beta', \mathbf{v}'$ should satisfy

$$\mathbf{v}' = \frac{\mathbf{X}^T\eta}{\lambda}, \|\mathbf{v}'\|_\infty \leq 1, \mathbf{v}'^T\beta' = \|\beta'\|_1$$

which is equivalent to

$$|\mathbf{x}_i^T \eta| \leq \lambda, i = 1, 2, \ldots, p. \tag{24}$$

Then we plug $\mathbf{v}' = \frac{\mathbf{X}^T \eta}{\lambda}$ and $\mathbf{v}'^T \beta' = \|\beta'\|_1$ into Eq. (23):

$$f_1(\beta') = \inf_{\beta} f_1(\beta) = -\eta^T \mathbf{X} \beta' + \lambda \big(\frac{\mathbf{X}^T \eta}{\lambda}\big)^T \beta' = 0 \tag{25}$$

Therefore, the optimum value of problem (21) is 0.

Next, let us consider problem (22). Denote the objective function of problem (22) as $f_2(\mathbf{z})$. Let us rewrite $f_2(\mathbf{z})$ as:

$$f_2(\mathbf{z}) = \frac{1}{2}(\|\mathbf{z} - \eta\|_2^2 - \|\eta\|_2^2) \tag{26}$$

Clearly,

$$\mathbf{z}' = \operatorname*{argmin}_{\mathbf{z}} f_2(\mathbf{z}) = \eta$$

and

$$\inf_{\mathbf{z}} f_2(\mathbf{z}) = -\frac{1}{2}\|\eta\|_2^2$$

Combining everything above, we get the dual problem:

$$\begin{aligned} \sup_{\eta} \quad & g(\eta) = \eta^T \mathbf{y} - \frac{1}{2}\|\eta\|_2^2 \\ \text{subject to} \quad & |\mathbf{x}_i^T \eta| \leq \lambda, \, i = 1, 2, \ldots, p \end{aligned} \tag{27}$$

which is equivalent to

$$\begin{aligned} \sup_{\eta} \quad & g(\eta) = \frac{1}{2}\|\mathbf{y}\|_2^2 - \frac{1}{2}\|\eta - \mathbf{y}\|_2^2 \\ \text{subject to} \quad & |\mathbf{x}_i^T \eta| \leq \lambda, \, i = 1, 2, \ldots, p \end{aligned} \tag{28}$$

By a simple re-scaling of the dual variables η, i.e., let $\theta = \frac{\eta}{\lambda}$, problem (28) transforms to:

$$\begin{aligned} \sup_{\theta} \quad & g(\theta) = \frac{1}{2}\|\mathbf{y}\|_2^2 - \frac{\lambda^2}{2}\|\theta - \frac{\mathbf{y}}{\lambda}\|_2^2 \\ \text{subject to} \quad & |\mathbf{x}_i^T \theta| \leq 1, \, i = 1, 2, \ldots, p \end{aligned} \tag{29}$$

A.2 Relationship Between The Primal And Dual Variables

Problem (18) is clearly convex and its constraints are all affine. By Slater's condition, as long as problem (18) is feasible we will have strong duality. Denote β^*, $\mathbf{z}^*$ and θ^* as optimal primal and dual variables. The Lagrangian is

$$L(\beta, \mathbf{z}, \theta) = \frac{1}{2}\|\mathbf{z}\|_2^2 + \lambda\|\beta\|_1 + \lambda\theta^T \cdot (\mathbf{y} - \mathbf{X}\beta - \mathbf{z}) \tag{30}$$

From the KKT condition, we have

$$0 \in \partial_\beta L(\beta^*, \mathbf{z}^*, \theta^*) = -\lambda \mathbf{X}^T \theta^* + \lambda \mathbf{v}, \text{ in which } \|\mathbf{v}\|_\infty \leq 1 \text{ and } \mathbf{v}^T \beta^* = \|\beta^*\|_1 \tag{31}$$

$$\nabla_{\mathbf{z}} L(\beta^*, \mathbf{z}^*, \theta^*) = \mathbf{z}^* - \lambda\theta^* = 0 \tag{32}$$

$$\nabla_{\theta} L(\beta^*, \mathbf{z}^*, \theta^*) = \lambda(\mathbf{y} - \mathbf{X}\beta^* - \mathbf{z}^*) = 0 \tag{33}$$

From Eq. (32) and (33), we have:

$$\mathbf{y} = \mathbf{X}\beta^* + \lambda\theta^* \tag{34}$$

From Eq. (31), we know there exists $\mathbf{v}^* \in \partial\|\beta^*\|_1$ such that

$$\mathbf{X}^T\theta^* = \mathbf{v}^*,\ \|\mathbf{v}^*\|_\infty \leq 1 \text{ and } (\mathbf{v}^*)^T\beta^* = \|\beta^*\|_1$$

which is equivalent to

$$|\mathbf{x}_i^T\theta^*| \leq 1, i = 1, 2, \ldots, p, \text{ and } (\theta^*)^T\mathbf{X}\beta^* = \|\beta^*\|_1 \tag{35}$$

From Eq. (35), it is easy to conclude:

$$(\theta^*)^T\mathbf{x}_i \in \begin{cases} \text{sign}(\beta)_i^* \text{ if } \beta_i^* \neq 0 \\ [-1, 1] \text{ if } \beta_i^* = 0 \end{cases} \tag{36}$$

B Deviation of the Dual Problem of Group Lasso

B.1 Dual Formulation

Assuming the data matrix is $\mathbf{X}_g \in \Re^{N\times n_g}$ and $p = \sum_{g=1}^G n_g$, the group Lasso problem is given by:

$$\inf_{\beta\in\Re^p} \frac{1}{2}\|\mathbf{y} - \sum_{g=1}^G \mathbf{X}_g\beta_g\|_2^2 + \lambda\sum_{g=1}^G \sqrt{n_g}\|\beta_g\|_2 \tag{37}$$

Let $\mathbf{z} = \mathbf{y} - \sum_{g=1}^G \mathbf{X}_g\beta_g$ and problem (37) becomes:

$$\begin{aligned} \inf_{\beta} \quad & \frac{1}{2}\|\mathbf{z}\|_2^2 + \lambda\sum_{g=1}^G \sqrt{n_g}\|\beta_g\|_2 \\ \text{subject to} \quad & \mathbf{z} = \mathbf{y} - \sum_{g=1}^G \mathbf{X}_g\beta_g \end{aligned} \tag{38}$$

By introducing the dual variables $\eta \in \Re^N$, the Lagrangian of problem (38) is:

$$L(\beta, \mathbf{z}, \eta) = \frac{1}{2}\|\mathbf{z}\|_2^2 + \lambda\sum_{g=1}^G \sqrt{n_g}\|\beta_g\|_2 + \eta^T \cdot (\mathbf{y} - \sum_{g=1}^G \mathbf{X}_g\beta_g - \mathbf{z}) \tag{39}$$

and the dual function $g(\eta)$ is:

$$g(\eta) = \inf_{\beta,\mathbf{z}} L(\beta, \mathbf{z}, \eta) = \eta^T\mathbf{y} + \inf_{\beta}\left(- \eta^T\sum_{g=1}^G \mathbf{X}_g\beta_g + \lambda\sum_{g=1}^G \sqrt{n_g}\|\beta_g\|_2\right) + \inf_{\mathbf{z}}\left(\frac{1}{2}\|\mathbf{z}\|_2^2 - \eta^T\mathbf{z}\right) \tag{40}$$

In order to get $g(\eta)$, let us solve the following two optimization problems.

$$\inf_{\beta} -\eta^T\sum_{g=1}^G \mathbf{X}_g\beta_g + \lambda\sum_{g=1}^G \sqrt{n_g}\|\beta_g\|_2 \tag{41}$$

and

$$\inf_{\mathbf{z}} \frac{1}{2}\|\mathbf{z}\|_2^2 - \eta^T\mathbf{z} \tag{42}$$

Let us first consider problem (41). Denote the objective function of problem (41) as

$$\hat{f}(\beta) = -\eta^T \sum_{g=1}^{G} \mathbf{X}_g \beta_g + \lambda \sum_{g=1}^{G} \sqrt{n_g} \|\beta_g\|_2 \tag{43}$$

Let

$$\hat{f}_g(\beta_g) = -\eta^T \mathbf{X}_g \beta_g + \lambda \sqrt{n_g} \|\beta_g\|_2, \qquad g = 1, 2, \ldots, G$$

then we can split problem (41) into a set of subproblems. Clearly $\hat{f}_g(\beta_g)$ is convex but not smooth because it has a singular point at 0. Consider the subgradient of $\hat{f}_g$,

$$\partial \hat{f}_g(\beta_g) = -\mathbf{X}_g^T \eta + \lambda \sqrt{n_g} \mathbf{v}_g, \qquad g = 1, 2, \ldots, G$$

where $\mathbf{v}_g$ is the subgradient of $\|\beta_g\|_2$:

$$\mathbf{v}_g \in \begin{cases} \frac{\beta_g}{\|\beta_g\|_2} & \text{if } \beta_g \neq 0 \\ \mathbf{u},\ \|\mathbf{u}\|_2 \leq 1 & \text{if } \beta_g = 0 \end{cases} \tag{44}$$

Let β_g' be the optimal solution of $\hat{f}_g$, then β_g' satisfy

$$\exists \mathbf{v}_g' \in \partial \|\beta_g'\|_2, \quad -\mathbf{X}_g^T \eta + \lambda \sqrt{n_g} \mathbf{v}_g' = 0.$$

If $\beta_g' = 0$, clearly, $\hat{f}_g(\beta_g') = 0$. Otherwise, since $\lambda \sqrt{n_g} \mathbf{v}_g' = \mathbf{X}_g^T \eta$ and $\mathbf{v}_g' = \frac{\beta_g'}{\|\beta_g'\|_2}$, we have

$$\hat{f}_g(\beta_g') = -\lambda \sqrt{n_g} \frac{(\beta_g')^T}{\|\beta_g'\|_2} \beta_g' + \lambda \sqrt{n_g} \|\beta_g'\|_2 = 0.$$

All together, we can conclude the

$$\inf_{\beta_g} \hat{f}_g(\beta_g) = 0, \quad g = 1, 2, \ldots, G$$

and thus

$$\inf_{\beta} \hat{f}(\beta) = \inf_{\beta} \sum_{g=1}^{G} \hat{f}_g(\beta_g) = \sum_{g=1}^{G} \inf_{\beta_g} \hat{f}_g(\beta_g) = 0.$$

The second equality is due to the fact that β_g's are independent.

Note, from Eq. (44), it is easy to see $\|\mathbf{v}_g\|_2 \leq 1$. Since $\lambda \sqrt{n_g} \mathbf{v}_g' = \mathbf{X}_g^T \eta$, we get a constraint on η, i.e., η should satisfy:

$$\|\mathbf{X}_g^T \eta\|_2 \leq \lambda \sqrt{n_g}, \qquad g = 1, 2, \ldots, G.$$

Next, let us consider problem (42). Since problem (42) is exactly the same as problem (22), we conclude:

$$\mathbf{z}' = \operatorname*{argmin}_{\mathbf{z}} \frac{1}{2} \|\mathbf{z}\|_2^2 - \eta^T \mathbf{z} = \eta$$

and

$$\inf_{\mathbf{z}} \frac{1}{2} \|\mathbf{z}\|_2^2 - \eta^T \mathbf{z} = -\frac{1}{2} \|\eta\|_2^2$$

Therefore the dual function $g(\eta)$ is:

$$g(\eta) = \eta^T \mathbf{y} - \frac{1}{2} \|\eta\|_2^2.$$

Combining everything above, we get the dual formulation of the group Lasso:

$$\begin{aligned} \sup_{\eta} \quad & g(\eta) = \eta^T \mathbf{y} - \frac{1}{2}\|\eta\|_2^2 \\ \text{subject to} \quad & \|\mathbf{X}_g^T \eta\|_2 \leq \lambda\sqrt{n_g},\ g = 1, 2, \ldots, G \end{aligned} \tag{45}$$

which is equivalent to

$$\begin{aligned} \sup_{\eta} \quad & g(\eta) = \frac{1}{2}\|\mathbf{y}\|_2^2 - \frac{1}{2}\|\eta - \mathbf{y}\|_2^2 \\ \text{subject to} \quad & \|\mathbf{X}_g^T \eta\|_2 \leq \lambda\sqrt{n_g},\ g = 1, 2, \ldots, G \end{aligned} \tag{46}$$

By a simple re-scaling of the dual variables η, i.e., let $\theta = \frac{\eta}{\lambda}$, problem (46) transforms to:

$$\begin{aligned} \sup_{\theta} \quad & g(\theta) = \frac{1}{2}\|\mathbf{y}\|_2^2 - \frac{\lambda^2}{2}\|\theta - \frac{\mathbf{y}}{\lambda}\|_2^2 \\ \text{subject to} \quad & \|\mathbf{X}_g^T \theta\|_2 \leq \sqrt{n_g},\ g = 1, 2, \ldots, G \end{aligned} \tag{47}$$

B.2 Relationship Between The Primal And Dual Variables

Clearly, problem (38) is convex and its constraints are all affine. By Slater's condition, as long as problem (38) is feasible we will have strong duality. Denote β^*, $\mathbf{z}^*$ and θ^* as optimal primal and dual variables. The Lagrangian is

$$L(\beta, \mathbf{z}, \theta) = \frac{1}{2}\|\mathbf{z}\|_2^2 + \lambda \sum_{g=1}^{G} \sqrt{n_g}\|\beta_g\|_2 + \lambda\theta^T \cdot (\mathbf{y} - \sum_{g=1}^{G} \mathbf{X}_g\beta_g - \mathbf{z}) \tag{48}$$

From the KKT condition, we have

$$0 \in \partial_{\beta_g} L(\beta^*, \mathbf{z}^*, \theta^*) = -\lambda \mathbf{X}_g^T \theta^* + \lambda\sqrt{n_g}\mathbf{v}_g, \text{ in which } \mathbf{v}_g \in \partial\|\beta_g^*\|_2, \quad g = 1, 2, \ldots, G \tag{49}$$

$$\nabla_{\mathbf{z}} L(\beta^*, \mathbf{z}^*, \theta^*) = \mathbf{z}^* - \lambda\theta^* = 0 \tag{50}$$

$$\nabla_{\theta} L(\beta^*, \mathbf{z}^*, \theta^*) = \lambda \cdot (\mathbf{y} - \sum_{g=1}^{G} \mathbf{X}_g\beta_g^* - \mathbf{z}^*) = 0 \tag{51}$$

From Eq. (50) and (51), we have:

$$\mathbf{y} = \sum_{g=1}^{G} \mathbf{X}_g\beta_g^* + \lambda\theta^* \tag{52}$$

From Eq. (49), we know there exists $\mathbf{v}_g' \in \partial\|\beta_g^*\|_2$ such that

$$\mathbf{X}_g^T \theta^* = \sqrt{n_g}\mathbf{v}_g'$$

and

$$\mathbf{v}_g' \in \begin{cases} \frac{\beta_g^*}{\|\beta_g^*\|_2} & \text{if } \beta_g^* \neq 0 \\ \mathbf{u},\ \|\mathbf{u}\|_2 \leq 1 & \text{if } \beta_g^* = 0 \end{cases}$$

Then the following holds:

$$\mathbf{X}_g^T \theta^* \in \begin{cases} \sqrt{n_g}\frac{\beta_g^*}{\|\beta_g^*\|_2} \text{ if } \beta_g^* \neq 0 \\ \sqrt{n_g}\mathbf{u},\ \|\mathbf{u}\|_2 \leq 1 \text{ if } \beta_g^* = 0 \end{cases} \tag{53}$$

for $g = 1, 2, \ldots, G$. Clearly, if $\|\mathbf{X}_g^T \theta^*\|_2 < \sqrt{n_g}$, we can conclude $\beta_g^* = 0$.